**ASSESSMENT IN MEDICAL EDUCATION: REVIEW AND CRITICAL APPRAISAL OF VARIOUS FORMATS**

**ABSTRACT**

**Background and Objectives:** The attention given to improving the art and science of teaching methods in medical education is increasing without corresponding attention to evaluation and assessment, which are the final key components of any learning process that aim to determine whether the set objectives of the learning process have been achieved or not. This review discusses various formats of assessment in clinical medicine, their merits and demerits, and how to improve the objectivity, validity, and reliability of various formats.

**Methods:** This discussion article was compiled based on observations and experiences of teachers of medical training over the years in assessing medical students and resident doctors. Also, relevant literature from internet were reviewed and findings from personal interviews of senior examiners in both undergraduate and postgraduate medical colleges were included.

**Conclusion:** There are various assessment formats and evaluations in medical education and using a single assessment format may not be sufficient to assess various components of the training thus careful selection and combinations of various formats should be employed based on the set objectives, requirements, and purpose of the assessment.

**Keywords:** medical education, examination, assessment format, review

**1.0 INTRODUCTION**

Tests and examinations are key components of any learning process to determine whether the set objectives of the learning process have been achieved or not.1 At some point, many medical practitioners will be involved in the act of assessing the medical student, residents or peers or other members of the healthcare team at one point or the other.

However, many medical practitioners and medical lecturers consider the act of testing and assessments as an act that comes naturally, just like the act of eating. This is a wrong perception as even the act of eating, if not properly learned, can lead to bad table manners. Many lecturers at various levels devote little time to this crucial aspect of education, which is the endpoint of any learning process, compared to time devoted to preparing and formulating lecture notes. Many doctors who are medical teachers also erroneously assume that the act of testing and evaluation comes by default.

Currently, in the United States, medical assessment is based on a model designed by the Accreditation Council for Graduate Medical Education (ACGME). This model assesses six domains of competence: knowledge, patient care, communication, professionalism, interpersonal skills, practice-based learning and improvement, and systems-based practice.2

Determining the most appropriate assessment format for a specific objective is difficult. This discussion article thus aimed to give a review and update on commonly used assessment formats highlighting their merits and demerits and briefly highlighting challenges associated with the assessment of clinicians' professional skills.

Based on our experience over the years as students who took various examinations at different levels, medical registrars, teachers, review of many previous examination questions, methods of evaluation, and interactions with specialists in test measurement and evaluation, we were compelled to write about the science of testing and evaluation in medical education. This review is not expected to exhaust all about the science of testing, measurements, and evaluation in medical education but is expected to remind the stakeholders of the need to make a conscious effort to employ basic scientific principles of the act of testing, measurement, and evaluation in medical education, an aspect of education often neglected and considered by many to come by default.

**2.0 METHODS**

The study was adopted a review design involving peer-reviewed articles focusing on assessment and evaluation in medical education. The following databases were used: Google Scholar, Pubmed, Scopus, Medline, and African Journal Online. Search words such as medical education or medicine training or residency training or medical assessment were used. Ninety-seven non-duplicate citations were screened of which 19 were excluded after the title and abstract were screened, seventy-eight were then retrieved for reading and after reading through, 70 articles were found to be relevant for the review process of this study.

Information was also obtained from commentaries, letters to the editor, and editorials. Articles that did not meet systemic research design and reporting were excluded. (Figure 1) We also interviewed experts in the field of science of testing, measurement, and evaluation, many medical examiners at various levels of medical education, and students from various medical schools. Authors’ observations and experience over the years in assessing medical students and resident doctors were also included. We then review and appraise the merits and demerits of various formats

Figure 1. Flow diagram of articles selection

**3.0 DISCUSSIONS**

**3.1 ASSESSMENT IN MEDICAL EDUCATION**

The primary objective of the examination is to assess the extent of knowledge gained by students against the set objectives at various levels of Bloom's taxonomy (Figure 1)3 based on the student level. The assessment is expected to serve the following functions:

* To motivate and inspire students for future learning.
* To discriminate between competent and incompetent physicians to be
* For the selection of an applicant who are likely to be successful in advanced training

Though assessment of knowledge is basic to any learning process however it is not enough in clinical medicine to determine the core competency and proficiency in various clinical skills which are of importance in medical practice. Thus, the learner needs to use such knowledge in various clinical scenarios.

Figure 2. Bloom’s taxonomy level of understanding

To achieve delivery of quality health care services there is a need to improve other components of assessment beyond knowledge level to include means of assessing other skills such as technical skills, communication skills, and interdisciplinary care through an evidence-based system. Thus, the assessment format should be tailored beyond knowledge level but be more encompassing to include assessment of other necessary skills. Thus, using Bloom’s taxonomy as a ladder paradigm for assessment in medical education will be insufficient to cater for other components of objectives of medical education such as the assessment of actual practical skills of the trainees in terms of competency and proficiency. This shortcoming of Bloom’s taxonomy can be overcome using Miller’s hierarchical model of assessment (Figure 2),4 which starts with an assessment of knowledge and ends with an assessment of behaviour and skills, thus making Miller’s model more appropriate than traditional Bloom’s taxonomy model as a paradigm for evaluation and assessment in medical education.



Figure 3. Miller’s hierarchical model for assessment of clinical competence

Based on Miller’s model, the prediction of professional competency and proficiency is closer as one moves up the Miller’s ladder.

The peculiarity of medical education was previously exemplified by a study that shows that theoretical knowledge gained in training has little predicting power of clinical skill and surgical competence among surgical residents.5 This could have been one of the reasons that many postgraduate colleges and many senior college elders believed in the need for yearly progress evaluations of trainees and the use of simulators in assessing the clinical skills of trainees. Also, to achieve assessment and evaluation of trainees in medical education, there is a great need to assess trainee's level of competencies, proficiencies, and performances in various aspects of the clinical skills. Thus, assessment in the medical examination should go beyond assessing trainees’ core theoretical knowledge.

Based on the recognition of the failure and inadequacy of evaluating various components of competency of training in medical examination through a single method of assessment, it is thus necessary to employ combinations of various means of assessment to assess trainees’ core competency in various aspects of clinical practice. Thus, the demerit of an assessment method can be compensated by other methods' merits. This thus served as the scientific basis for using combinations of various formats of assessments at both formative and summative levels of examination based on the set objectives.

**3.2 FORMS OF ASSESSMENT**

**Formative** **assessment**: this is usually in training assessment aimed primarily at guiding future learning, providing reassurance, shaping values, and encouraging reflection. It assists the students who are in the early phase of learning to appreciate their deficiencies and thus serves to motivate them to set higher goal standards 6 They are usually used in low stakes examinations.

**Summative assessment:** this form of assessment is usually used at the end of training for a decision to make an overall judgment about a candidate's competence, ability to practice, or qualification for promotion to a higher level of practice or role.

**3.3 FORMATS OF ASSESSMENT IN MEDICAL EXAMINATIONS**

All forms of assessment format have merits and intrinsic demerits. However, the combination of different formats can help to compensate for such demerits.

Assessment in medical education falls into six main domains which are knowledge, patient care, communication, professionalism, interpersonal skills, practice-based learning and improvement, and systems-based practice.2

 Assessments of knowledge can be in the form of a free response format (FRF) (SAQ, SEQ, LAQ) or a selected response format (SRF) (MCQ and its variants) while skill performance assessment takes the form of clinical examination in various formats such as the traditional clinical clerkship of long and short cases and Objective Structured Clinical Examination (OSCE) and picture test model of assessment. 7

The assessment format(s) employed in assessing trainees will depend on the aim and objectives of the examination, whether for theoretical knowledge or competence/proficiency in a particular set of clinical skills, or application of knowledge in a clinical setting. However, no single method of assessment can satisfactorily be considered ideal over the other as each method has its own merits and demerits. Irrespective of the method used, the quality/usefulness of the assessment method can be assessed using the following basic criteria:

* **Content validity**: This assesses whether the test assesses what is expected i.e., is the content appropriate for the level?
* **Reliability**: This assesses whether the test consistently measures what is expected to be measured. This simply means that if a candidate passes a particular test in a certain area, he should be expected to pass a similar test in the same area and vice versa if failed. However, if the area tested is too narrow (sample error) the score obtained may not be reliable enough to decide on the whole subject.
* **Clarity**: is the question framed without being ambiguous?
* **Discriminatory power**: That is, the ability of a test to discriminate between good and poor examinees with high fidelity.
* **Educational impact**: this determines the effect an assessment has on future learning and practice.
* **Acceptability to learners and teachers**:
* **Cost**: the overall cost of the assessment format to learner, teacher, and the institution

**3.4 ASSESSMENT OF KNOWLEDGE**

**3.4.1 WRITTEN EXAMINATIONS**

**3.4.1.1 LONG ANSWER QUESTIONS** **(LAQ)**

This format of assessment is a written examination that is designed to assess a trainee’s level of knowledge and application of such knowledge in a specific aspect of training and how the trainees can organize their thoughts. It tests the ability of the trainee to process, evaluate, synthesize, and organize information to conclude by summarizing their thoughts over a specified time. This format of assessment is considered more suitable when trainees are expected to explain or discuss a specific clinical case. When framing an essay question, it is expected that the question should be clear and devoid of ambiguity and that a marking scheme should have been formulated before marking.

This format of assessment is heralded with some inherent defects such as:

**Poor scope**: the scope of LAQ is narrowed and thus sampling error is high. This can be exemplified in a situation whereby a candidate engaged in the study of a pattern of previous examination questions and geared his/her reading toward some recurring set of questions. Such a candidate may avoid reading other aspects of the course and still pass the examination if their permutation analysis is correct.

**The bias of judgment in marking**: this often results from variation in the judgment of the markers/examiners, opinion of the marker as well as mood and state of mind of the marker; for example, a marker who just got promoted may be elated and score candidates differently from a marker who has just been denied of promotion. Also, a marker may be biased based on the marker’s area of interest. Consider a situation where a question is framed as “discuss colorectal carcinoma.” A clinical surgeon marking scheme may be biased towards surgical management while an academic surgeon may be biased towards the molecular aspect and a surgical oncologist may lay more emphasis on medical (chemotherapeutic) management.

**Candidate’s factors**: this may include legibility of writing/handwriting, spelling errors, choice of words, and arrangement of the write-up that may not necessarily represent a candidate’s depth of knowledge.

The shortcomings of essay-type questions can be minimized by assigning multiple markers to mark a single question (double marking) and compare scores or a single marker marking a set of questions thus reducing variational bias.8

**3.4.1.2 SHORT ESSAY QUESTIONS (SAQ)**

This format of examination tends to minimize some deficits of LAQ as it tends to cover a larger scope with short responses. However, it is not devoid of other deficits of LAQ, and all the measures to minimize the deficit of LAQ are also applicable.

**3.4.1.3 MODIFIED ESSAY QUESTIONS**

This is a modification of traditional essay questions that tries to test the problem-solving ability of a candidate through reasoning9 rather than operating at Bloom’s taxonomy knowledge level by supplying a rigid sequential answer to a clinical scenario. It relies on question interdependency in which failure to answer a preceding question correctly will lead to a wrong response to subsequent questions. The candidate also does not have the opportunity to review or go back to previous questions.

**3.4.2 KEY FEATURES QUESTIONS**

Key feature questions try to test a specific aspect of a candidate’s clinical decision ability in a specific clinical case.10 These questions can take various formats such as multiple-choice or open-ended questions. Because of its validity and reliability, it has been suggested to be used when testing candidate clinical decision skill ability; and when properly formulated with adherence to the guidelines.11 Formulating key feature questions consumes time, especially in situations where the examiners are not conversant with this format of examination.12 it can be used to test a candidate's ability to apply knowledge to solve a clinical case in a high-stakes examination.

Table 1: Sample question

|  |
| --- |
| **A sample feature key question**A 55-year-old woman presented with an eczematous skin lesion on her right breast which starts from the nipple and spreads to the areolar region. Your next line of action will be.1. Commence broad spectrum antibiotic therapy
2. Commence application of topical steroid-based therapy
3. Reassure the patient and watch
4. Refer the patient to a dermatologist
5. Considered for tissue biopsy and histology
 |

**2.4.3 MULTIPLE CHOICE QUESTIONS (MCQ)**

This format of examination is the most common form of assessment in medical examination. The format covers a broad area and practically addresses many shortcomings of LAQ and SAQ when well-constructed. MCQ takes various forms; the form in which the candidate chooses the best option out of 4 to 5 options and true or false format. Table 1

**Forms of MCQ**

1. True or false type
2. Five-choice completion with the best answer option
3. Five-choice negative form completion type
4. Five or four-choice association type
5. Combination type
6. Case history analysis type
7. Diagram, tracing, or photographic type
8. Relationship analysis there are two sentences; a statement and a reason for the statement

MCQ has a very high reliability per unit time of testing because it covers a wider scope, thus reducing sampling errors. MCQ has the benefit of being more objective, covering a wider scope, being easy to mark manually or electronically, and having answers that are not equivocal. 13 However, many erroneously consider MCQ tests to be limited to assessing only level 1 of Bloom’s taxonomy. When properly constructed, they can be used to test the application of knowledge and a candidate’s ability to solve problems in a clinical scenario.14 MCQ can be framed to be context-free when the primary aim is to test basic knowledge and fact findings, or it can be contextualized by framing a clinical scenario that requires a decision based on weighing and analysis of both clinical and laboratory parameters.15 MCQ can also give more extra information on an examination such as no or poor teaching effect, and can also help to detect examinees percentile more objectively in an examination and make a better comparison of student academic performance of two different institutions. 16Though, MCQ allows guessing, however, if the best-in-five type of model is employed, the chance that a guesser will pass in a set of 100 questions is 1 in 500 and 1 in 1000 if the number of questions is increased to 200. Setting an MCQ is more time-consuming and taxing for the examiner when aiming for a quality test. However, when the total time taken for reading and marking essay type of questions is compared, it takes almost the same time to set good-quality MCQ questions. Also, the MCQ format has the demerit of cueing effect in which it allows the examinee to pick an answer correctly by mere recognition of the option which could not have been gotten when such a question is open-ended without options. 17 18

The most important factor when setting good quality MCQ is to include effective, reasonable distractors and bait options. The bait can be an option that will directly relate to the questions while distractors will resemble the correct answer and often serve as a challenge to the examinee and must be carefully picked to serve the function. There is a need to review the bank of MCQ questions at various intervals, and the questions should be pooled from different specialties and ensure that 10 to 15 aspects of the subject matter are tested for good coverage. During a review of MCQs, irrelevant, ambiguous, and controversial answers should be avoided, and questions with the use of “always” and “never” should be excluded.19 The response to correct options should be spread across options A, B, C, D, E, and not skewed to a particular option alphabet. Further information and guidelines about writing such questions can be obtained from Case and Swanson.16,20

**2.4.4 SCRIPT CONCORDANCE TEST (SCT) FORMAT SCT**

The format is designed to test a candidate’s thought and reasoning process in ambiguous situations21 and is currently gaining more acceptance as an assessment format in medical education. The candidate will be given clinical scenarios and then presented with a set of related questions in three sections. The first section will be phrased as “if you are thinking of” a set of relevant diagnoses or management options; the second section will be phrased as “if you are to find” a set of clinical findings, and the third section will be framed as “your option would become” a choice from a 5-point Likert scale.

The clinical scenario below illustrates a sample of the SCT format construct: A 20-year-old female known sickle cell patient presented with abdominal pain and inability to pass faeces for 5 days before presentation. You entertained a clinical diagnosis of acute abdomen secondary to acute appendicitis.

Table 2: Clinical scenario and hypothesis development based on respondent’s agreement

|  |  |  |  |
| --- | --- | --- | --- |
| **If you are thinking of (a situation)** | **If you are to find that (additional information)** |  | **This hypothesis would become (extent of agreement)** |
| Ordering for AXR | No abdominal tenderness  |  | SA A N DA SD |
| Ordering for D-dimer | Unilateral recent unilateral calf swelling |  | SA A N DA SD |
| Ordering for PT | On her menses  |  | SA A N DA SD |
| Epigastric pain | NSAID abuse | Acute gastritis  | SA A N DA SD |

Investigative action new findings level of agreement based on Likert scale.

The candidate option will then be compared in concordance with the judgment made by the panel of reference experts.22 An SCT allows testing and assessing candidates for real-life situations that are difficult to assess by other formats. A high degree of concordance between the candidate and the panel of expert reference is considered optimal and appropriate and can be considered to measure the quality of clinical reasoning.23 SCT validity and reliability have been documented in various specialities. Though SCT is often used in clinical settings, it could be used in ethical or professional issues.24-28 For more information regarding this format the reader is referred to Demester and Charlin's article. 29

**3.5 PERFORMANCE AND CLINICAL SKILL ASSESSMENT**

Assessment competency and proficiency in clinical skills is one of the hallmarks of medical examination and its peculiarity. The importance of clinical skill assessment is reflected in the gravity and effect of failure in this component of the examination in which the current and undisputable agreement is that failure in this component of the examination translates to final failure irrespective of the candidate's overall score 30,31 and such failure is often termed *clinical failure* to emphasis its importance.

Performance assessment is done through clinical examinations in various formats. This form of assessment is considered the most important in the medical examination as it is what a clinician will encounter throughout their career in the medical profession.

**3.5.1 LONG CASES**

This is the traditional way of assessing clinical acumen in medical training. In this format, the examinee is given a specific amount of time to interact with a real patient and then presents their findings to a set of examiners who will assess the examinee through an unstructured question and by dictate of flow during the discussion of the case. The long case has face validity and authenticity since it mimics a real situation. A long case is subjective as its outcome may be affected by the examiner’s perception of the case among other factors. Also, some examinees may have difficult patients with ambiguous clinical cases while other examinees may have simple cases. In scoring examinees in long cases, the score should be divided into various components of the presentation and not just the final diagnosis.

The various components should include the structure of the presentation, eliciting of relevant physical signs, and discussion. Of note in recent times is the surprising pattern of clinical presentation among examinees and doctors in which during physical examination, many examinees will preferentially start physical examination from a suspected region of pathology. This form of practice should be discouraged as the primary aim of long cases is beyond making a diagnosis of the patient’s primary complaint but the discovery of other relevant pathology that the patient may not even volunteer. In many instances, such practice makes the examinee downgrade the importance of a detailed clinical examination, thus missing such hidden pathology. Such erroneous practice was once seen in a doctor who only examined a patient’s abdomen because of gross abdominal distention in a patient with endomyocardial fibrosis, and we asked him where the primary pathology was, in which case the primary pathology was in the heart. Also, we saw a doctor who devoted most of his examination to the lower limb examination in a patient with unilateral lower limb swelling, however, the primary pathology was a rectal cancer with pressure effect on the pelvic lymphatic with lymphatic infiltration.

In essence, the organ of primary pathology cannot be defined or known until a full thorough and detailed systematic examination from head to toe is done and is preferably done sequentially so as not to be biased during such examination. Also, a review of the system should not be neglected in the long case as it is the only opportunity to detect other hidden pathology that may affect a patient’s subsequent plan.

The use of long cases in high-stakes summative examinations should be discouraged as it has poor reliability due to its subjectivity and poor scope.32 However, it was considered to have a good educational impact when used as a formative assessment for a low-stakes examination. 30 33

**3.5.2 SHORT CASE**

This format of the examination is designed to assess the candidate's clinical competence in eliciting physical signs.34  In which the examinee will be asked to examine a real patient in the presence of two or more examiners that will award a score based on his attitude towards the patient, technical skill in demonstrating certain physical signs and a short discussion on the case. The examinee is expected to see at least two cases and the more cases an examinee sees the more the score is likely to be awarded.

**3.5.3 MINI CLINICAL EVALUATION EXERCISES (MINI-CEX)**

Mini-CEX was developed by the American Board of Internal Medicine with the primary aim of assessing six core clinical competencies that include medical interviewing skills, physical examination skills, humanistic qualities/professionalism, clinical judgment, counselling skills, organization, and efficiency through observation of the trainee by the trainers during their day-to-day encounter.35The observed performance is scored using a Likert-like scale of 1 to 4; with 1 corresponding to unacceptable, 2 to below expectation, 3 to meet expectation, and 4 to exceeded expectation. The observation and score can be done many times by different trainers thus allowing the opportunity to identify a missed component of core competency a trainee failed to exhibit during training. This format is more suitable for formative assessment by direct trainers of the trainees.36

**3.5.4 ORAL EXAMINATION**

Oral examination often called viva voce in which a set of two or three panels of examiners will assess examinees knowledge in various aspects of clinical care. The section also includes an assessment of examinees in anatomic pathology, identification of instruments, and various aspects of the application of laboratory medicine to clinical scenarios. The section also features an assessment of examinees on medical and surgical emergencies. However, this section of the examination is not devoid of subjectivity as a familiar candidate may be asked simpler questions while others will be asked a more relatively difficult question thus making this method have higher inter-rater variability. This subjectivity is currently being minimized in many high-stakes examinations using a selection of questions from a standardized pool of questions whereby a candidate will be asked to pick a question number from a certain range or pick a question blindly from a pool of questions boxes. The assessment in this part of the examination may also be fairer as the format allows the examiners to discuss and analyze examinees' responses before scoring.

**3.5.5 OBJECTIVE STRUCTURED CLINICAL EXAMINATION (OSCE**)

OSCE was introduced in the mid-seventies as a form of assessment aimed at assessing basic clinical skills in medical education37 and is rapidly replacing the traditional clinical examination in many medical examinations globally due to the perceived objectivity in its assessment. In this format of the examination, the candidates will move from one station to another station with each station set up to assess a particular clinical situation such as an examination of the abdomen under the watch of the examiner who will grade the candidate or a candidate may be asked to take a short or focused history on particular clinical symptoms while in some instance a station may focus on a short clinical history and laboratory report which the candidate will ask to study and this will be followed up by series of questions. About 18 to 30 stations are recommended for better reliability and coverage.38

The examination is conducted by using a simulator that would have been trained to mimic a clinical condition or use real or standardized patients. The main benefits of this format include the ability to examine many candidates at the same time, the candidates are likely to cover a larger spectrum of clinical conditions in a short time and all candidates will be exposed to the same set of questions thus minimizing discrepancy in difficulty level across candidates. However, the format is associated with poor assessment of the candidate’s depth of knowledge. The format may also not represent true-life situations as in most cases the simulator responses are often stereotyped towards a perfect clinical situation, for example, a patient with a distended abdomen will be made to give four classical symptoms of intestinal obstruction when a clinical scenario is to depict intestinal obstruction which may not reflect a real clinical scenario a as medicine and human pathology are is not an exact science as often depicted in many clinical OSCE scenarios. To improve the objectivity of OSCE use of checklists for scoring are often recommended and it is important to avoid the inclusion of not-too-relevant points in checklists to maintain their validity.39 When developing a checklist, it is also recommended that experts in the field should set and agree on the component and scoring rules of the checklist to ensure content validity and reliability.40

**3.6 IN-TRAINING RECORD (PORTFOLIO) ASSESSMENT FORMAT**

This format of assessment used in-training records that include the use of documented evidence of trainees' work over certain periods this may include the use of logbooks, collections of case reports, and records of participation in certain procedures and activities among others. Also, trainees are expected to have periodic documentation of reflection learning experiences. Assessment of in training record will involve many trainers that the trainee passed through, and this is done using checklists or periodic global ratings based on required standards. This form of assessment is commonly employed in formative assessment in many medical schools and is less suitable for summative assessment41, 42 Guidelines for this format can be found in Thistle Thwaite43andFriedman et al.44

**3.7 ASSESSING BEHAVIOURAL SKILLS IN MEDICAL EXAMINATION**

The medical school graduates and newly qualified fellows need to possess both cognitive and non-cognitive skills to function professionally and provide quality medical services that go beyond cognitive knowledge.31

[The current mode of examination in many medical schools and postgraduate examinations may not be reliable in predicting leadership and judgment skill ability to function as a good physician following exiting examinations.](https://bmcmededuc.biomedcentral.com/articles/10.1186/s12909-024-05046-5) 45 Thus, the use of an assessment format that possesses the ability to predict such personal and professional qualities is important in medical school and postgraduate fellowship examinations.

Multiple Mini Interviews (MMIs) are currently employed to assess such qualities in medical school admissions.

**3.7.1 MULTIPLE MINI INTERVIEWS MMIs**

This format was introduced at McMaster University in 2004 to provide solutions to previously identified gaps in current assessment formats, which are: (i) Failure of traditional interview (oral examination) to reliably predict medical school performance. (ii) Patient complaints that require non-cognitive skills such as judgment, communication, ethics, and empathy cannot be reliably addressed by current traditional means of assessment.

MMI assessment is designed in a way that examinees are exposed to 6 to 10 clinical situations over a time. Each situation is specifically designed to assess a predetermined specific non-cognitive skill such as the ability to empathize, communication skills, and leadership skills such as the examinee’s ability to counsel and take consent, resolve an ethical issue or dilemma within the stipulated time under the watch of two examiners who will independently score the examinee. [MMIs have been shown to have high reliability, feasibility, and good predictive validity in predicting professional quality](https://bmcmededuc.biomedcentral.com/articles/10.1186/s12909-024-05046-5).46 The format also can assess the candidate’s ability to work under pressure and in a team. 46

**3.8 MULTISOURCE OR 360O ASSESSMENTS**

The clinical practice involves interaction with various members of staff in hospitals and patients, thus assessments by multiple assessors such as peers, nurses, administrative staff, supervisor and patients coupled with trainee self-assessment can help to predict the trainee's ability to communicate with other team members and the extent of interpersonal relationships47-49 Peer rating has been shown by some studies to be reliable50 and insightful,51,52 and comparable to supervisors rating.53 Peer assessments when lacking trust and confidence may be destructive. Patient rating is heralded with some problems such as the severity of illness often affects rating compliance and rating score as very ill patients tend to have low compliance and tend to rate clinicians low compared to a mildly ill patient,54 also the need for a high number of patients to rate for reliability. 55 Ratings by nurses are more reliable, 56 and often correlate with other team members' ratings. 57

This form of assessment is better in formative assessment and rather used for developmental evaluation rather than for the determination of passing or failing an examination. 58 A sample form to use in 360o evaluations can be obtained from the Wood et al study. 59

**3.9 ASSESSING PROCEDURAL SKILL**

**Direct Observation of Procedural Skills**. This is a format of assessment in which a faculty member observes the trainee while performing a specific procedure such as the setting of an IV-line, passage of urethral catheter, aseptic technique, endoscopy procedures to surgical procedures with real patients, or with the use of manikin. This format of assessment is essential for assessing core competence and proficiency in clinical skills. As for clinical skills, direct observation is essential for the assessment of procedural skills. This format can also be used to assess trainee interpersonal skills such as counselling and communication skills with patients.60,61 This form of assessment can be exemplified when a faculty member assesses a trainee while assisting with surgery or while the faculty member plays the role of assistance to the trainee who serves as the lead surgeon during a surgical procedure.

**3.10 GUIDE FOR CHOOSING ASSESSMENT FORMAT**

There is gamut of assessment formats available to choose from however the format or combinations of formats to be chosen should be based on some factors such as the aim of the assessment, the level of medical education, stakes of the examination whether high stakes or low stakes examination. For example, high-stakes summative assessment should be more comprehensive and involve combinations of many formats because the consequence and implication of the outcome are very important as many of such assessments are exiting assessments.

Summative in-training assessment is better designed to measure discrete components of knowledge or skills as there are enough time to assess various components of training. Also, such a method of summative assessment will allow early recognition if training fails to advance and progress in a specific component of training as compared to when multiple integrated assessments are done at the same time which may allow a candidate to compensate from another component of training resulting in final graduation of an unbalancing candidate.

For knows and knows how (knowledge, application of Bloom’s taxonomy) of Miller’s pyramid of clinical assessment essay type format and MCQ should be considered, while for show how of Miller’s pyramid (clinical competency) OSCE, should be considered, and for those of Miller’s pyramid (performance assessment) MMI, portfolio and use of logbooks should be considered.

**3.11 MARKING SYSTEM**

There are two forms of marking essay questions and clinical components of the examination, namely closed and open marking systems.

**3.11.1 CLOSED MARKING SYSTEM**

This form of marking system is the traditional system of marking in many medical examinations. This system of marking brings most grades to the middle of the score line in which no candidate is given a score beyond a certain set point and not below a certain set point. Based on the scoring principle, it has poor discriminatory power and tends to put high-ranking candidates at a disadvantage, especially when such a candidate transcript is required for admission into other programs.

However, there is a recent argument for a paradigm shift to the use of a modified closed marking system or an expanded closed marking system. A closed marking system minimizes the effect of the subjective nature of the score by individual assessors when marking essay type of examination. It also prevents score compensation from a deficit in a particular section of the examination where a candidate scores so low. This system of scoring may be better suitable for postgraduate fellowship examination as the fellowship examination is not designed to rank the examinees in order of their performance but rather to determine whether the examinees meet the set of criteria and aim to produce a balanced candidate rather than a candidate with narrow knowledge in one area and broad knowledge in another area.

**3.11.2 OPEN MARKING SYSTEM**

In this marking system, the examinee can score any mark based on the content of the write-up, and the scores are not limited to a certain range as in a closed marking system. However, the effect of subjectivity in score may be more exaggerated as previously highlighted under the demerit of essay type of assessment. However, this effect can be reduced using double assessors who will score the examinee based on an agreed structured marking scheme designed by the court of examiners, and when a candidate’s score is at variance with a score from another section of the examination as candidate script should be considered for further review by the senior examiner and another independent marker.

**3.12 CONDUCT OF EXAMINERS DURING EXAMINATION**

During the conduct of examination, especially the clinical component of the examination, no candidate is expected to be assisted. However, a candidate can be guided in clinical or oral examinations, and if the candidate decides not to take the clue the examiner need not harass the candidate. The act of nodding head by some examiners when a candidate is on the right path should be discouraged. So also, the gestural act of disagreement when a candidate is wrong should be discouraged during the conduct of the examination. Verbal utterances such as “yes,” “good,” “okay,” and any other positive or negative comment a candidate can interpret as signs of agreement or disagreement respectively should all be discouraged during the conduct of the examination. The examinee should be allowed to talk without unnecessary interruption, as such an act may not only interfere with the candidate’s flow of thought and organization of the presentation but may convert the examination session to another lecture session and waste the candidate’s time and deprive the candidate of the required score. Also, this form of interruption by the lead examiner should be avoided as much as possible to allow co-examiners to have their own time to assess the candidate as such an act of neglect may irritate the co-examiners.62,63 For further information on conduct during examinations, the reader is referred to O.G Ajao's article.63

**4.0 CONCLUSION**

There are various formats of assessment and evaluation in clinical medicine and using a single format of assessment may not be sufficient to assess various components of the training due to limitations and strengths associated with each format. Therefore careful selection and combinations of various formats as discussed should be employed based on the set objectives, requirement, and purpose of the assessment whether for low stakes or high stakes summative or formative examination.

**REFERENCES**

1. Zaidi NL, Kreiter CD, Castaneda PR, et al. Generalizability of competency assessment scores across and within clerkships: How students, assessors, and clerkships matter. *Acad Med.*2018;93:1212–1217.
2. Batalden P, Leach D, Swing S, Dreyfus H, Dreyfus S. General competencies and accreditation in graduate medical education. Health Aff (Millwood) 2002;21:103-111
3. Bloom BS. *Handbook I: Cognitive domain.* New York: David McKay; 1956. Taxonomy of educational objectives.
4. Miller GE. The assessment of clinical skills/ competence/performance. *Acad Med.*1990;65:S63–7.
5. Cowles JC. Hubbard JF. A comparative study of essay and objective examinations for medical students. J Med Edu 1952; 27(2):14.
6. Friedman Ben-David M. The role of assessment in expanding professional horizons. Med Teach 2000;22:472-477new
7. Schuwirth LW, van der Vleuten CP.  Different written assessment methods: what can be said about their strengths and weaknesses? *Med Educ.*2004;38:974–979.
8. Schuwirth LWT, van der Vleuten CP. In: *Written Assessments.* Dent J, Harden R, editors. New York: Elsevier Churchill Livingstone; 2005. pp. 311–22.
9. Knox JD. How to use modified essay questions. *Med Teach.*1980;2:20–4.
10. Bordage G, Page G. In: *An alternate approach to PMPs, the key feature concept. Further developments in assessing clinical competence.* Hart I, Harden R, editors. Montreal: Can-Heal Publications; 1987. pp. 57–75.
11. Farmer E, Page G. A practical guide to assessing clinical decision-making skills using the key features approach. *Med Educ.*2005;39:1188–94.
12. Schuwirth LWT, van der Vleuten CP. ABC of learning and teaching in medicine: Written assessment. *BMJ.*2003;326:643–5.
13. Mahmood H. Correlation of MCQ and SEQ scores in written undergraduate ophthalmology assessment. *J Coll Physicians Surg.*2015;25:185–188.
14. Hift RJ. Should essays and other "open-ended"-type questions retain a place in written summative assessment in clinical medicine? BMC Med Educ. 2014 Nov 28;14:249. doi: 10.1186/s12909-014-0249-2.
15. Schuwirth LW, Verheggen MM, van der Vleuten CP, Boshuizen HP, Dinant GJ. Do short cases elicit different thinking processes than factual knowledge questions do? *Med Educ.*2001;35:348–56.
16. Hubbard JP and Clemens WV. Multiple choice questions in medicine. Philadelphia: Lea and Febiger. 1961.
17. Schuwirth LW, van der Vleuten CP, Donkers HH. A closer look at cueing effects in multiple-choice questions. Med Educ 1996;30:44-49
18. Schuwirth LW, van der Vleuten CP. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ 2004;38:974-979
19. Ajao OG. Multiple choice question for specialist examination in General surgery. Ibadan University Press. 1982.
20. Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences. From <http://www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf>. [Accessed December 2023].
21. Charlin B, van der Vleuten CP. Standardized assessment of reasoning in context of uncertainty. The Script Concordance Test approach. *Eval Health Profess.*2004;27:304–19.
22. Hall KH. Reviewing intuitive decision-making and uncertainty: the implications for medical education. *Med Educ.*2002;36:216–24.
23. Charlin B, Vleuten C Van Der. Standardized assessment of reasoning in context of uncertainty. The Script Concordance Test approach. *Evaluation and the Health Professions.*2004;**27**:304–319. doi: 10.1177/0163278704267043.
24. Gagnon R, Charlin B, Coletti M, Sauvé E, Vleuten C Van der. Assessment in the context of uncertainty: How many members are needed on the panel of reference of a script concordance test? *Medical Education.*2005;**39**:284–291. doi: 10.1111/j.1365-2929.2005.02092.x.
25. Charlin B, Gagnon R, Pelletier J, Coletti M, Abi-Rizk G, Nasr C, Sauvé E, Vleuten C Van der. Assessment of clinical reasoning in the context of uncertainty: the effect of variability within the reference panel. *Medical Education.*2006;**18**:22–27.
26. Demeester A. *Mémoire de maîtrise.* Vol. 13. Bobigny – Université de Paris; 2004. Évaluation des processus de Raisonnement clinique des étudiants sages-femmes en fin de cursus.
27. Meterissian S, Zabolotny B, Gagnon R, Charlin B. Is the script concordance test a valid instrument for assessment of intraoperative decision-making skills? *American Journal of Surgery.*2007;**193**:248–51. doi: 10.1016/j.amjsurg.2006.10.012.
28. Gagnon R, Charlin B, Roy L, Saint Martin M, Sauvé E, Boshuizen HPA, Vleuten C Van der. The cognitive validity of the script concordance test: a processing time study. *Teaching and Learning in Medicine.*2006;**18**:22–27. doi: 10.1207/s15328015tlm1801\_6.
29. Fournier JP, Demeester A, Charlin B. Script Concordance Tests: Guidelines for Construction. *BMC Med Inform Decis Mak.*2008;8:18.
30. Stokes JF. Medical Education Booklet. No 2 Dundee Association for the study of medical education; 1974. The Clinical; examination.
31. Ajao OG, Alao AO. Objective Structured Clinical Examination (OSCE) *J of Surgery and Surgical Sciences.*2008;1:1–4.
32. Ponnamperuma GG, Karunathilake IM, McAleer S, Davis MH. The long case and its modifications: a literature review. *Med Educ.*2009;43:936–41
33. Wass V, van der Vleuten C. The long case. *Med Educ.*2004;38:1176–80.
34. Wass V, van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *Lancet.*2001;357:945–9.
35. Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: a method for assessing clinical skills. *Annals Internal Med.*2003;138:476–83.
36. Rees C, Shepherd M. The acceptability of 360-degree judgements as a method of assessing undergraduate medical students’ personal and professional behaviours. *Med Educ.*2005;39:49–57.
37. Harden RM, Gleeson FA. ASME Medical Education Booklet No. 8: Assessment of clinical competence using an objective structured clinical examination (OSCE) *Med Educ.*1979;13:41–54.
38. ACGME Outcome Project, Accreditation Council for Graduate Medical Education (ACGME) and American Board of Medical Specialist (ABMS) Toolbox of assessment methods. Sep, 2000. Version 1.1 From <https://www.acgme.org/globalassets/PDFs/Milestones/Guidebooks/AssessmentGuidebook.pdf> [Accessed January 2024].
39. Van der Vleuten CP, Schuwirth LW. Assessing professional competence: from methods to programmes. *Med Educ.*2005;39:309–17.
40. Marks M, Humphrey-Murto S. In: *Performance Assessment.* Dent J, Harden R, editors. New York: Elsevier Churchill Livingstone; 2005. pp. 323–35.
41. Rees C, Sheard C. The reliability of assessment criteria for undergraduate medical students’ communication skills portfolios: The Nottingham experience. *Med Educ.*2004;38:138–44.
42. Davis M, Friedman B, Harden R, Howie P, McGhee C, Pippard M, et al. Portfolio assessment in medical students’ final examinations. *Med Teach.*2001;23:357–66.
43. Thistlethwaite J. How to keep a portfolio. *Clin Teach.*2006;3:118–23.
44. Friedman MB, David MH, Harden RM, Howie PW, Ker J, Pippard MJ. AMEE guide No. 24: Portfolios as a method of student assessment. *Med Teach.*2001;23:535–51.
45. Yusoff MSB. Multiple Mini interviews as an admission tool in higher education: insights from a systematic review. J Taibah Univ Med Sci. 2019;14:203–40
46. 2022 FACTS: Applicants and Matriculants Data. 2022. At https://www.aamc.org/data-reports/students-residents/ data/2022-facts-applicants-and-matriculants-data.).
47. [Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. JAMA 1993;269:1655-1660](http://scholar.google.com/scholar_lookup?hl=en&publication_year=1997&author=C+Violato&author=A+Marini&author=J+Toews&author=J+Lockyer&author=H+Fidler&title=Feasibility+and+psychometric+properties+of+using+peers%2C+consulting+physicians%2C+co-workers%2C+and+patients+to+assess+physicians.)
48. Dannefer EF, Henson LC, Bierer SB, et al. Peer assessment of professional competence. Med Educ 2005;39:713-722
49. Violato C, Marini A, Toews J, Lockyer J, Fidler H. Feasibility and psychometric properties of using peers, consulting physicians, co-workers, and patients to assess physicians. Acad Med 1997;72:Suppl 1:S82-S84.
50. Lurie SJ, Nofziger AC, Meldrum S, Mooney C, Epstein RM. Temporal and group-related trends in peer assessment amongst medical students. Med Educ 2006;40:840-847.
51. Norcini JJ. Peer assessment of competence. Med Educ 2003;37:539-543.
52. Nofziger AC, Davis B, Naumburg EH, Epstein RM. The impact of peer assessment on professional development. Presented at the Ottawa Conference on Medical Education and Assessment, Ottawa, July 15, 2002.
53. Lurie S, Lambert D, Grady-Weliky TA. Relationship between dean's letter rankings, peer assessment during medical school, and ratings by internship directors. [Submitted for Publication] (in press).
54. Hall JA, Milburn MA, Roter DL, Daltroy LH. Why are sicker patients less satisfied with their medical care? Tests of two explanatory models. Health Psychol 1998;17:70-7.
55. Calhoun JG, Woolliscroft JO, Hockman EM, Wolf FM, Davis WK. Evaluating medical student clinical skill performance: relationships among self, peer, and expert ratings. Proc Annu Conf Res Med Educ 1984;23:205-210.
56. Butterfield PS, Mazzaferri EL. A new rating form for use by nurses in assessing residents' humanistic behavior. J Gen Intern Med 1991;6:155-161.
57. Klessig J, Robbins AS, Wieland D, Rubenstein L. Evaluating humanistic attributes of internal medicine residents. J Gen Intern Med 1989;4:514-52.
58. Haggins RS, Bridges J, Burke JM, O’Donnell MA, Cohen NM, Wilkes SB. Implementing the ACGME general competencies in a cardiothoracic surgery residency program using 360-degree feedback. *Annals Thoracic Surg.*2004;77:12–17.
59. Wood J, Collins J, Burnside ES, Albanese MA, Propeck PA, Kelcz F, et al. Patient, faculty, and self-assessment of radiology resident performance: a 360-egree method of measuring professionalism and interpersonal/ communication skills. *Acad Radiol.*2004;11:931–9
60. Husk KE, Learman LA, Field C, Connolly A. 2020. “Implementation and Initial Construct Validity Evidence of a Tool, myTIPreport, for Interactive Workplace Feedback on ACGME Milestones”. Journal of Surgical Education. Published online June 13, 2020. doi: 10.1016/j.jsurg.2020.05.00
61. Abdelsattar, J.M., Y.N. Aljamal, R.K. Ruparel, P.G. Rowse, S.F. Heller, and D.R. Farley. 2018. “Correlation of Objective Data with General Surgery Resident In-Training Evaluations and Operative Volumes.” Journal of Surgical Education 75: 1430-6. doi: 10.1016/j.jsurg.2018.04.016
62. Van der Vleuten,, VanLuyk SJ, Van Ballegooijen,, Swanson DB. Training and experience of examiners. *Med Educ.*1989;23:290–296.
63. Ajao OG and Ugwu BT. Evaluating the evaluators in postgraduate fellowship examinations - a viewpoint. [J West Afr Coll Surg.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4170259/) 2011;1(2): 97–104.